CUNY SPS DATA 621 - CTG5 - HW5

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

May 15th, 2019

Contents

Table 1: Data Dictionary

VARIABLE	DEFINITION	TYPE
TARGET	Number of Cases Purchased	count response
AcidIndex	Method of testing total acidity by using a weighted avg	continuous numerical predictor
Alcohol	Alcohol Content	continuous numerical predictor
Chlorides	Chloride content of wine	continuous numerical predictor
CitricAcid	Citric Acid Content	continuous numerical predictor
Density	Density of Wine	continuous numerical predictor
FixedAcidity	Fixed Acidity of Wine	continuous numerical predictor
FreeSulfurDioxide	Sulfur Dioxide content of wine	continuous numerical predictor
LabelAppeal	Marketing Score indicating the appeal of label design	categorical predictor
ResidualSugar	Residual Sugar of wine	continuous numerical predictor
STARS	Wine rating by a team of experts. 4 = Excellent, 1 = Poor	categorical predictor
Sulphates	Sulfate conten of wine	continuous numerical predictor
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	continuous numerical predictor
VolatileAcidity	Volatile Acid content of wine	continuous numerical predictor
pH	pH of wine	continuous numerical predictor

1 DATA EXPLORATION

When dining in a restaurant, a sommelier can assist you in selecting a perfect wine, even if you do not know much about wine yourself. By asking about your taste preferences, they can recommend a wine that pairs well with your meal, while complementing your likes and dislikes. But what happens when you're a large wine manufacturer, wondering how to produce wines that will sell? If the wine manufacturer can predict the number of cases sold based on the characteristics of a wine, then that manufacturer will be able to adjust their wine offering to maximize sales.

Our data set contains information on approximately 12,000 commercially available wines. Most of the variables are related to the chemical properties of the wine. The response variable is the number of sample cases that were purchased by wine distribution companies after sampling the wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States and promote further sales.

1.1 Summary Statistics

Continuous quantitative and categorical variables were summarized separately for the sake of clarity.

`STARS` and `LabelAppeal` can be described as ordinal categorical variables. However, because of the numerical coding, these variables were imported as if they were quantitative. Since they are ordinal consideration was given to treating them as numerical, but ultiately the decision was made to convert them to factors.

1.1.1 Summary Statistics Graphs

The histograms in Figure 1 shows the distribution of all 15 variables. The distribution is more peaked than a normal bell curve for most of our varibales especially `Chlorides`, `CitricAcid`, `Density`, `FixedAcidity` , `FreeSulfurDioxide`, `pH`, `ResidualSugar`, `Sulphates`, `TotalSulfurDioxide` and `ViolatileAcidity`. Most of those are also centered at or near zero. `AcidIndex` has a slight right skew.

Table 2: Summary statistics for numeric variables

	n	min	mean	median	max	sd
AcidIndex	12795	4.00	7.77	8.00	17.0	1.32
Alcohol	12142	-4.70	10.49	10.40	26.5	3.73
Density	12795	0.89	0.99	0.99	1.1	0.03
Sulphates	11585	-3.13	0.53	0.50	4.2	0.93
pH	12400	0.48	3.21	3.20	6.1	0.68
TotalSulfurDioxide	12113	-823.00	120.71	123.00	1057.0	231.91
FreeSulfurDioxide	12148	-555.00	30.85	30.00	623.0	148.71
Chlorides	12157	-1.17	0.05	0.05	1.4	0.32
ResidualSugar	12179	-127.80	5.42	3.90	141.2	33.75
CitricAcid	12795	-3.24	0.31	0.31	3.9	0.86
VolatileAcidity	12795	-2.79	0.32	0.28	3.7	0.78
FixedAcidity	12795	-18.10	7.08	6.90	34.4	6.32
TARGET	12795	0.00	3.03	3.00	8.0	1.93

Table 3: Summary statistics for categorical variables

	STARS	LabelAppeal
	1 :3042	-2: 504
	2 :3570	-1:3136
	3 :2212	0 :5617
	4 : 612	1 :3048
	NA's:3359	2 : 490

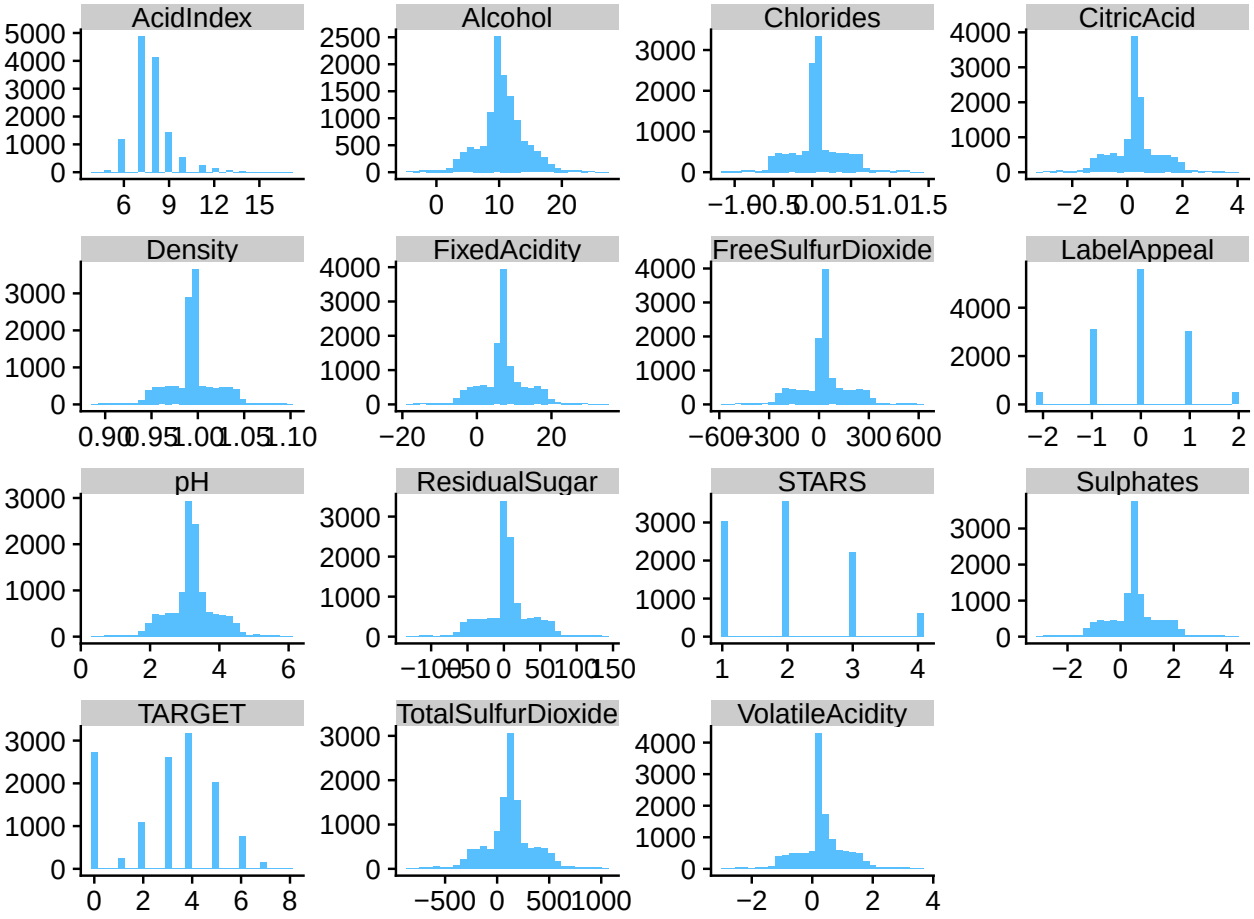


Figure 1: Data Distributions

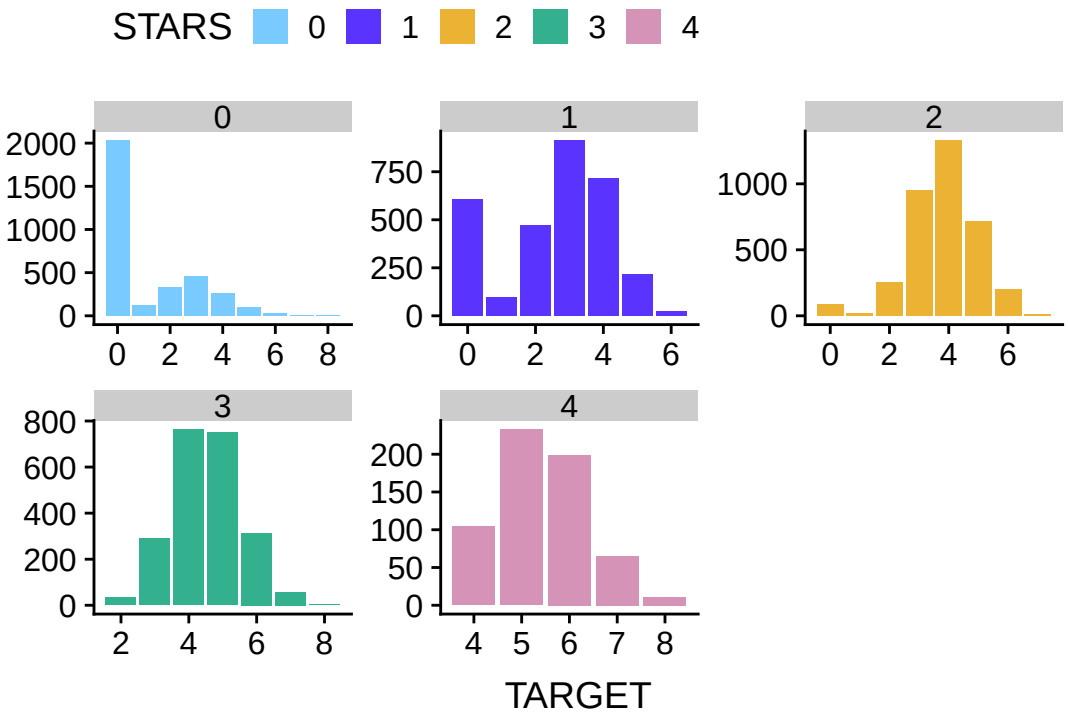


Figure 2: TARGET Distributions by STARS Values

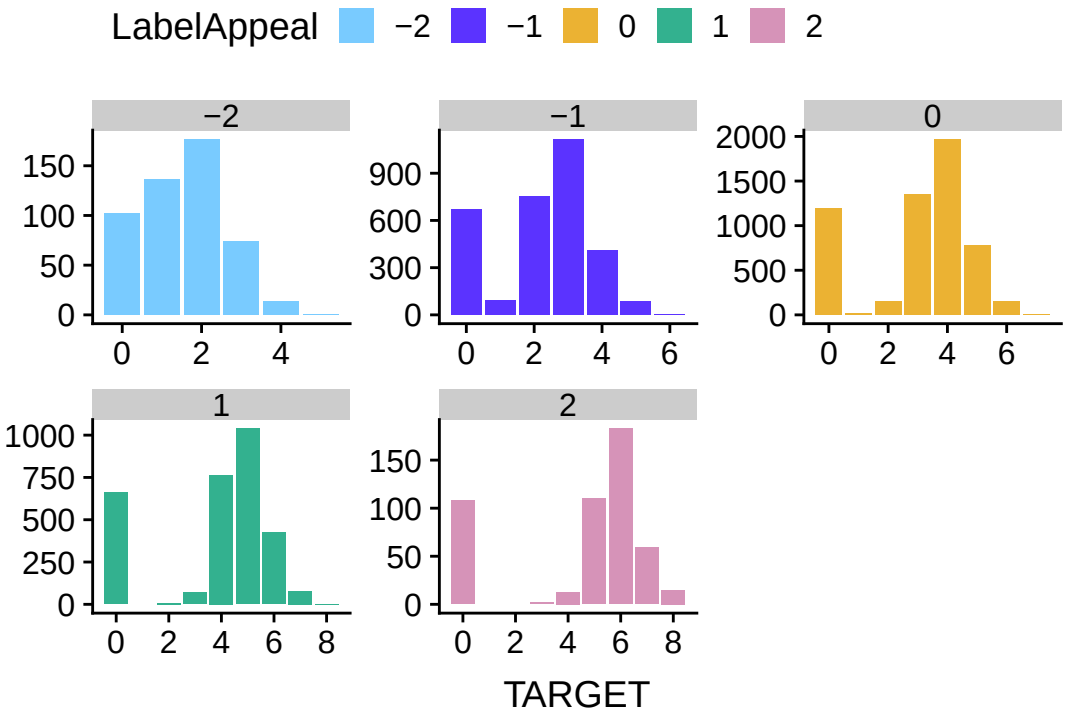


Figure 3: TARGET Distributions by LabelAppeal Values

Figure 2 shows that there are a large number of outliers that need to be accounted for, except for `LabelAppeal`, `AcidIndex` and `STARS` which have limited number of variations.

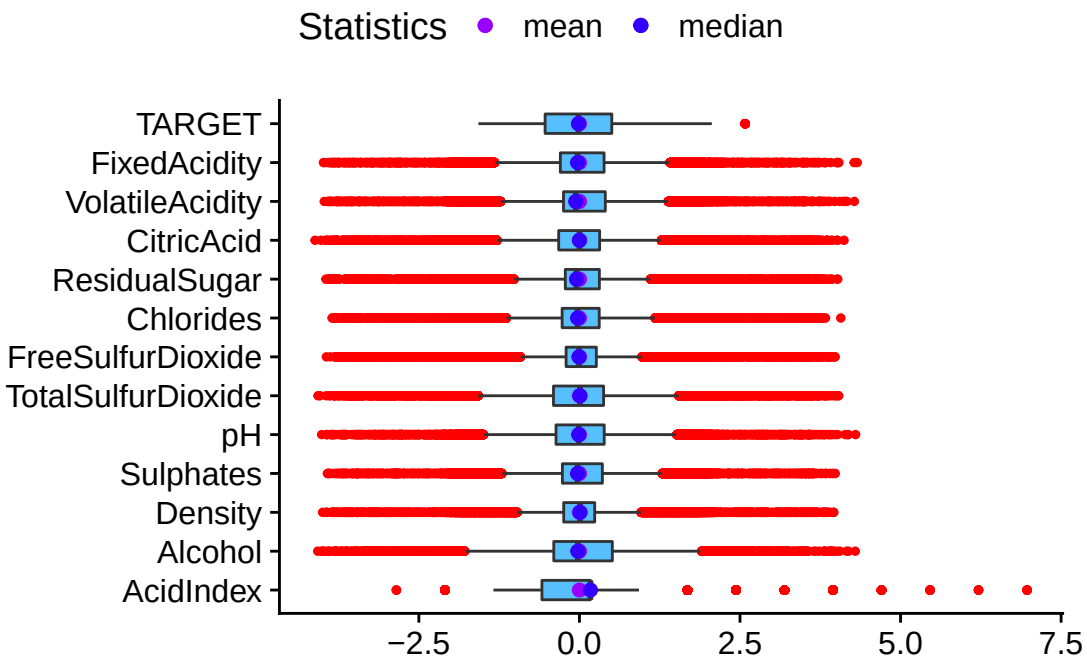


Figure 4: Scaled Boxplots

1.2 Linearity

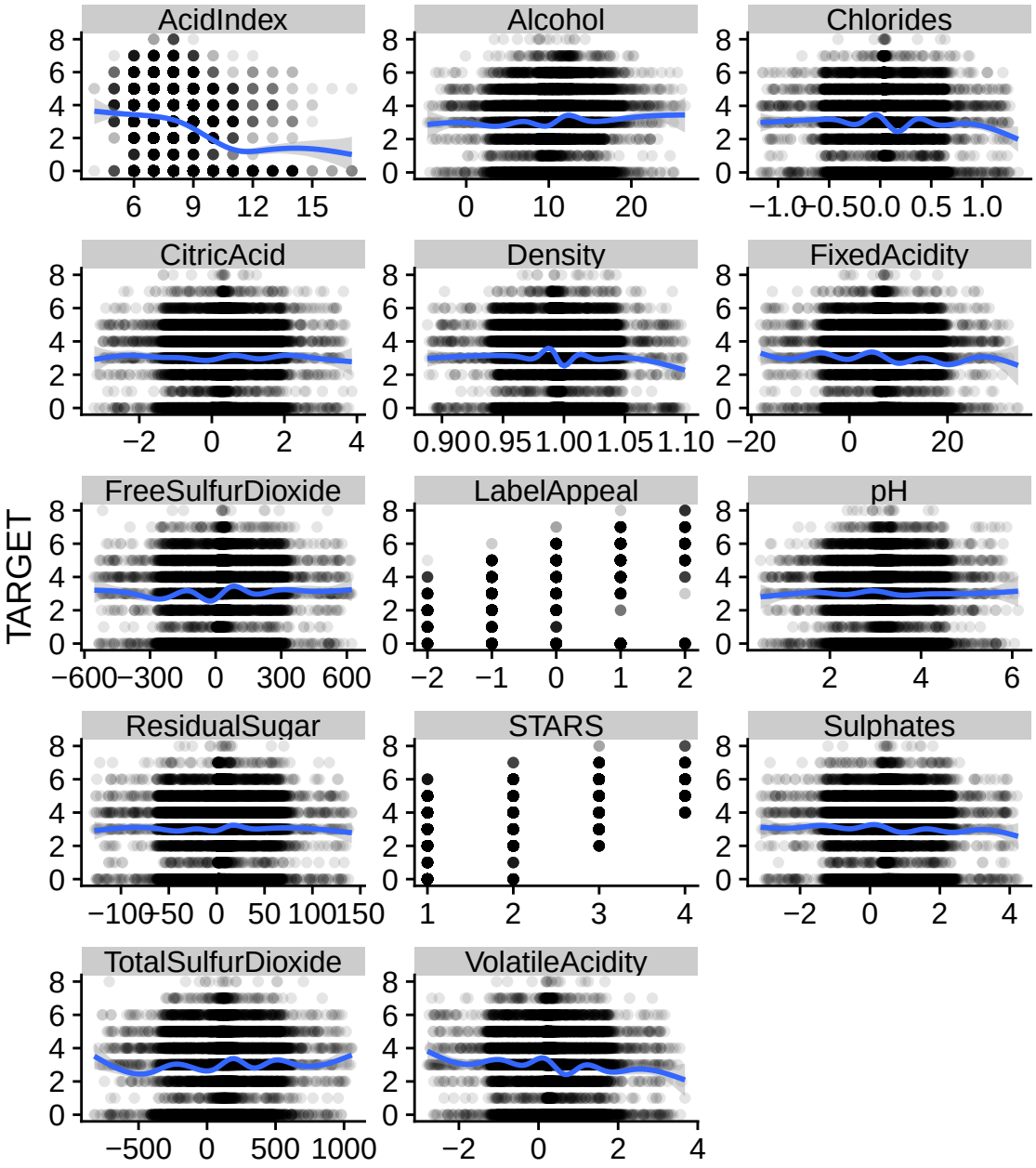


Figure 5: Scatter plot between numeric predictors and the TARGET

The raw predictors fail to show linear relationship with the `TARGET` except for the `AcidIndex` and `VolativeAcidity`. The Scatter Plots show a systematic, wave-like pattern for `Density`, `FixedAcidity`, `FreeSulfurDioxide`, `TotalSulfurDioxide` and `VolativeAcidity`.

1.2.1 Log Transformed Data

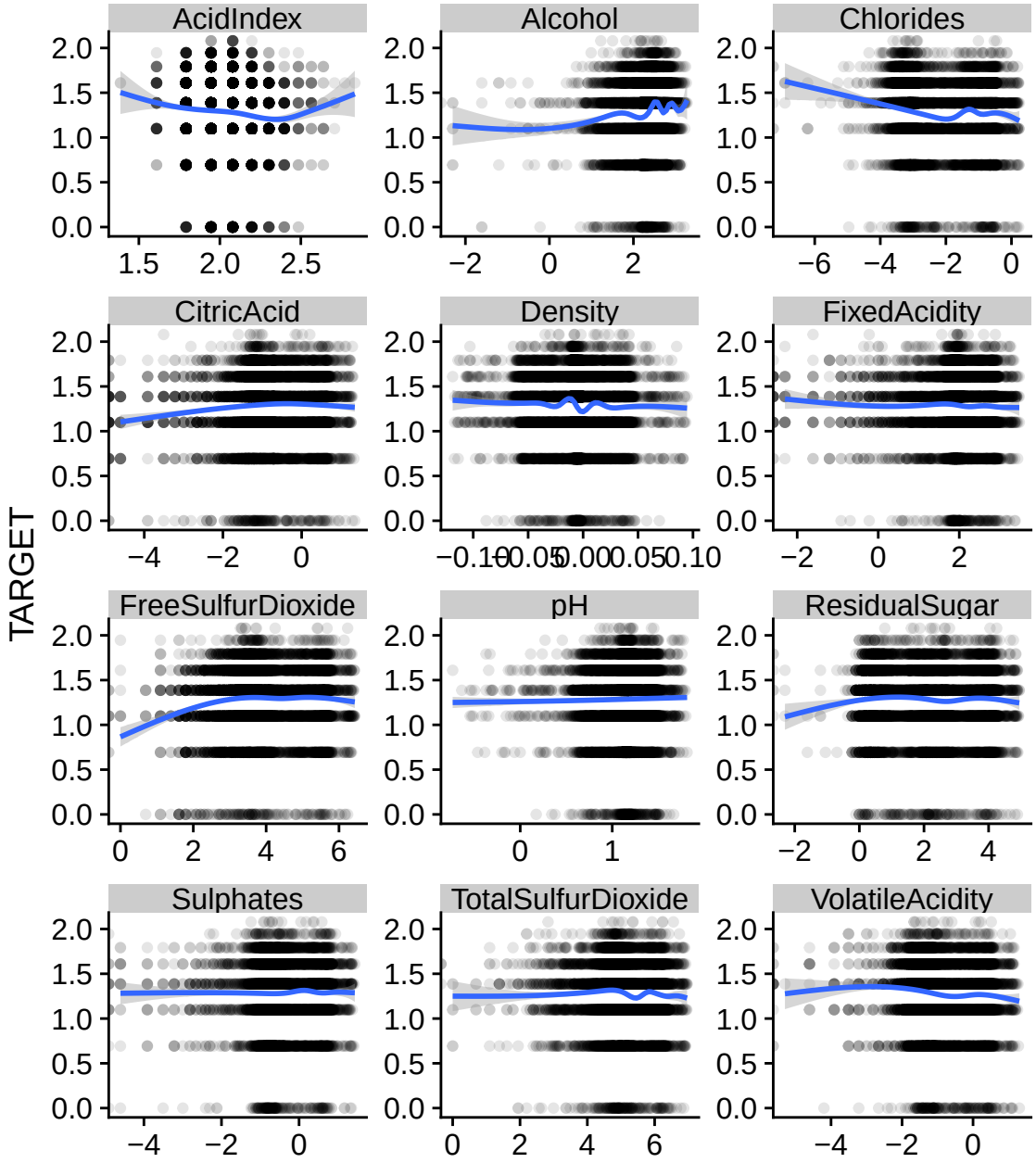


Figure 6: Scatter plot between log transformed predictors and the log transformed TARGET filtered for rows where TARGET is greater than 0

In attempt to improve the linearity of the variables against the `TARGET` variable, we start with a log transformation on all predictors and`TARGET` variable. As a result, the linearity of `Chlorides` and `FreeSulfurDioxide` become more apparent.

1.2.2 Box-Cox

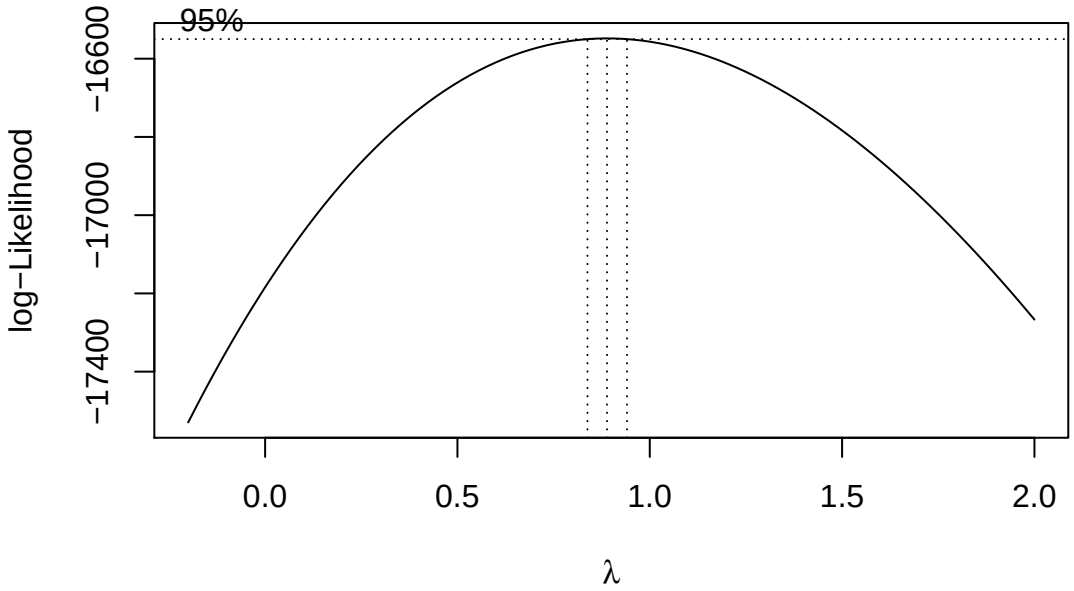


Figure 7: Box-Cox Plot

1.2.3 Square Root Transformed Predictors and Log Transformed Target

In 'Linear Models with R', Faraway suggested that the square root transformation is often appropriate for count response data. The Poisson distribution is a good model for counts, and that distribution has the property that the mean is equal to the variance thus suggesting the square root transformation. A plot of each predictor square root transformed plotted against the log transformed `TARGET`.

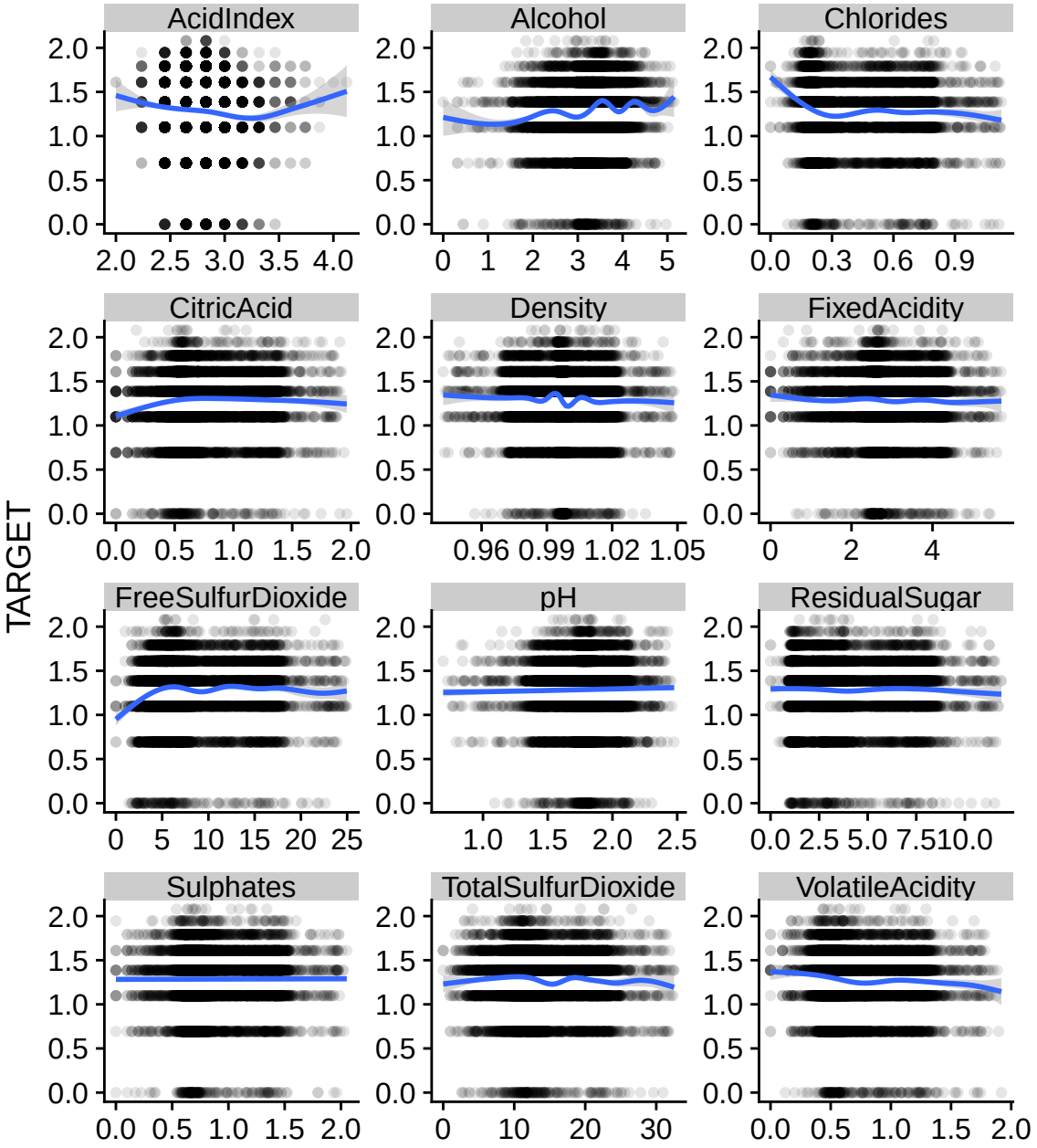


Figure 8: Scatter plot between square root transformed predictors and the square root transformed TARGET filtered for rows where TARGET is greater than 0

1.3 Missing Data

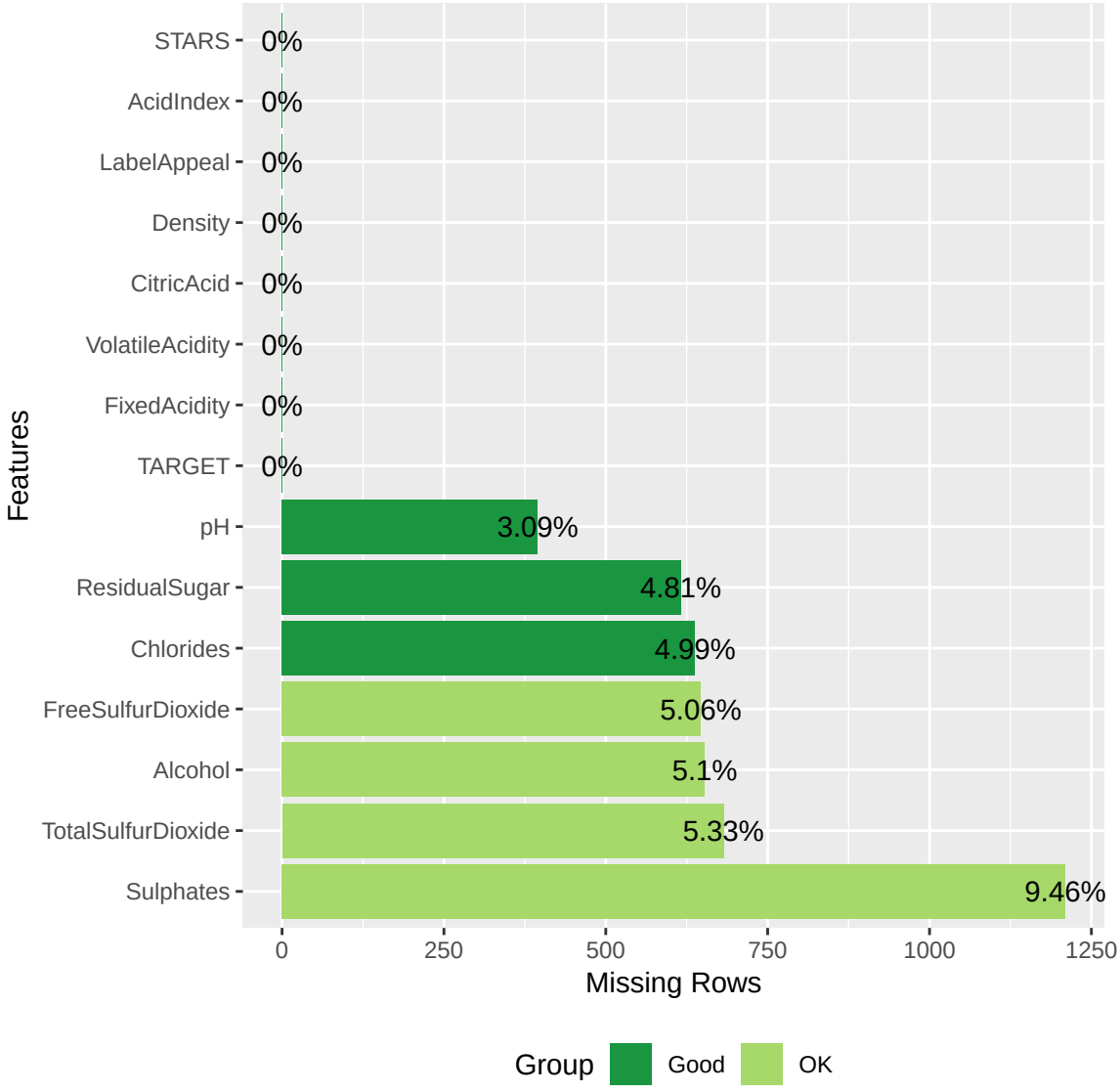


Figure 9: Missing data

A number of variables are missing observations: `STARS`, `Sulphates`, `TotalSulfurDioxide`, `Alcohol`, `FreeSulfurDioxide`, `Chlorides`, `ResidualSugar`, `pH`. For `STARS`, the number is 26.25%, but the others range between 3% and 9% of total. Approximately 50% of the cases are missing one of these variables.

2 DATA PREPARATION

Describe how you have transformed the data by changing the original variables or creating new variables

a. Fix missing values (maybe with a Mean or Median value)

b. Create flags to suggest if a variable was missing

c. Transform data by putting it into buckets

d. Mathematical transforms such as log or square root (or use Box-Cox)

e. Combine variables (such as ratios or adding or multiplying) to create new variables

[JO: Consider strategies to transform negative variables - arithmetic?]

2.1 Missing Values

[Rough explanation] About half of the cases are missing at least one of the following variables: `STARS`, `Sulphates`, `TotalSulfurDioxide`, `Alcohol`, `FreeSulfurDioxide`, `Chlorides`, `ResidualSugar`, or `pH`. We use MICE (Multivariate Imputation by Chained Equations) to impute values these values based on ... [JO following up]

Pink and blue lines indicate close fit match in distribution of imputed values and recorded values, with the exception of STARS - the addition of STARS values has given the appearance of shifting the distribution from high to low lambda values.

2.2 Transformation / Feature Engineering

There are a large number of negative values for variables for which that is nonsensical, examples: `Alcohol`, `CitricAcid`, `FixedAcidity`, `FreeSulfurDioxide`, `ResidualSugar`, `Sulphates`, `TotalSulfurDioxide`. and `VolatileAcidity`. The range for the Poisson and negative binomial distribution has zero as a lower bound, so we can arithmetically transform the aforementioned variables to scale the lower IQR non-outlier values from zero up and drop the sub-IQR values (now the only negative values remaining). [JO currently function not working as intended - moving all points up by the mean - IQR * 1.5, and tossing anything less than that - so tuning calculations]

(https://www.imachordata.com/do-not-log-transform-count-data-bitches/)

Alternatively, we can explore whether more information on how measurements were made can be found to discern whether there might be some reason for negative values, and if there's a possibility of systematic data errors

We'll also test the data for over-dispersion when setting up negative binomial models. [CODING UNDERWAY]

3 BUILD MODELS

Using the training data set, build at least two different poisson regression models, at least two diffe

Discuss the coefficients in the models, do they make sense? In this case, about the only thing you can

4 Poisson regression models

BR I am going to add both models under one heading...

5 Negative binomial regression models

5.1 Negative binomial regression model 1

5.2 Negative binomial regression model 2

6 Linear regression models

6.1 Linear regression model 1

6.2 Linear regression model 2

Table 4: Comparison of models

model	BIC	Log-likelihood	Degrees of freedom
Poisson-Logit Hurdle Regression	40806.810578686	-20275.7436342663	27
Zero-inflated Poisson	40900.671316578	-20322.6740032122	27
Negative Binomial 1	45781.5412721341	-22786.7457288325	23
Negative Binomial 2	45772.2839490919	-22782.1170673114	22
Linear Model	43336.4446687359	-21616.2098807507	11
Linear Model 2	42836.2580131354	-21186.4371677271	49

7 SELECT MODELS

7.1 Comparison of models

We will look at six different models, two poisson models, two negative-binomial models and an ols regression thrown in for good measure.

Both the Poisson-Logit Hurdle Regression and the zero-inflated poisson are very close in log likelihoods and BIC's.

The Poisson-Logit Hurdle Regression provides a closer fit to the observed than does the other models. The hurdle model is a modified count model in which there are two processes, one generating the zeros and one generating the positive values.

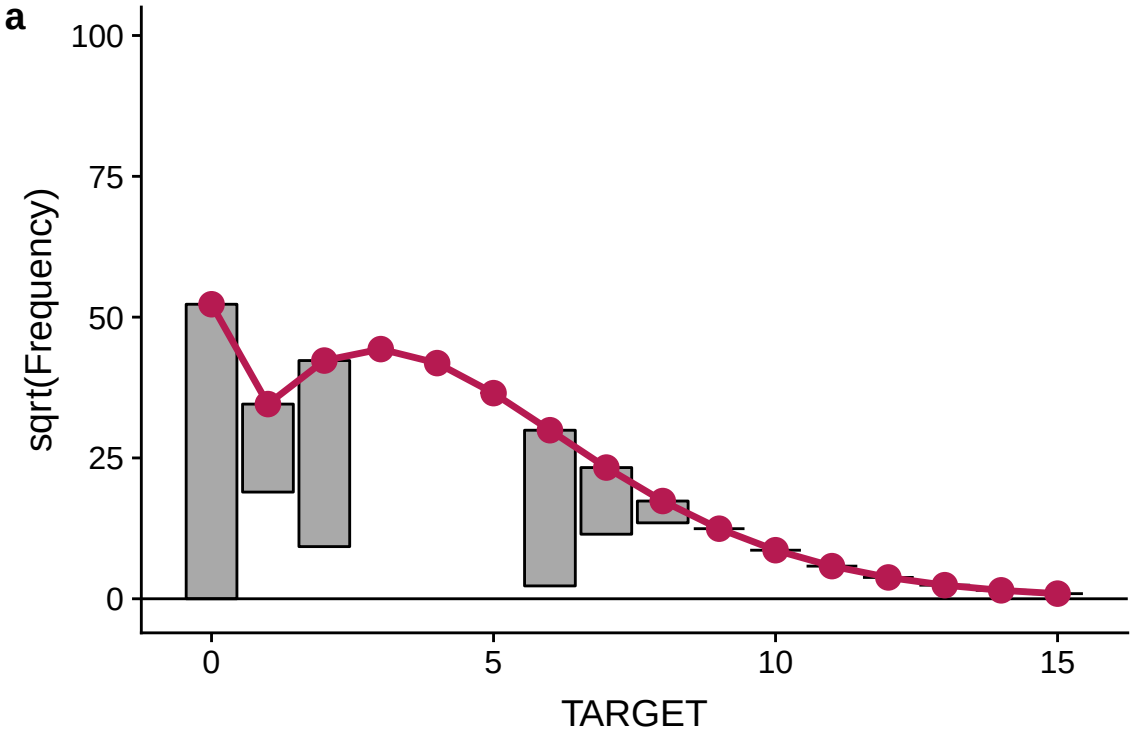
The Poisson-logit hurdle model is clearly the best choice here. The results for this model are given below.

```
##
## Call:
## hurdle(formula = TARGET ~ AcidIndex + Alcohol + LabelAppeal + STARS |
##     VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + pH +
##         Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS, data = minusinfluential)
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.13058 -0.44224 -0.00351  0.39520  4.55994
##
## Count model coefficients (truncated poisson with log link):
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept)    0.33498    0.06716    4.99           0.00000061 ***
## AcidIndex     -0.01706    0.00492   -3.47              0.00053 ***
## Alcohol        0.00750    0.00144    5.19           0.00000021 ***
## LabelAppeal-1  0.54013    0.04973   10.86 < 0.0000000000000002 ***
## LabelAppeal0   0.84353    0.04880   17.28 < 0.0000000000000002 ***
## LabelAppeal1   1.04033    0.04937   21.07 < 0.0000000000000002 ***
## LabelAppeal2   1.19964    0.05324   22.53 < 0.0000000000000002 ***
## STARS1         0.05401    0.02146    2.52              0.01186 *
## STARS2         0.16894    0.02002    8.44 < 0.0000000000000002 ***
## STARS3         0.25943    0.02097   12.37 < 0.0000000000000002 ***
## STARS4         0.36360    0.02592   14.03 < 0.0000000000000002 ***
## Zero hurdle model coefficients (binomial with logit link):
##                      Estimate Std. Error z value             Pr(>|z|)
## (Intercept)          4.563693   0.399257   11.43 < 0.0000000000000002 ***
## VolatileAcidity     -0.184953   0.036523   -5.06       0.000000410629 ***
## FreeSulfurDioxide    0.000645   0.000196    3.28              0.00103 **
## TotalSulfurDioxide   0.000827   0.000127    6.52       0.000000000069 ***
```

```
## pH                  -0.189424   0.042003   -4.51       0.000006491117 ***
## Sulphates           -0.117837   0.030709   -3.84              0.00012 ***
## Alcohol             -0.017513   0.007718   -2.27              0.02326 *
## LabelAppeal-1       -0.484425   0.137301   -3.53              0.00042 ***
## LabelAppeal0        -0.904500   0.134029   -6.75       0.000000000015 ***
## LabelAppeal1        -1.454734   0.143620  -10.13 < 0.0000000000000002 ***
## LabelAppeal2        -1.872201   0.222901   -8.40 < 0.0000000000000002 ***
## AcidIndex           -0.387795   0.021434  -18.09 < 0.0000000000000002 ***
## STARS1               1.828269   0.061392   29.78 < 0.0000000000000002 ***
## STARS2               4.270374   0.117280   36.41 < 0.0000000000000002 ***
## STARS3              20.247583 363.102538    0.06              0.95553
## STARS4              20.407039 694.552828    0.03              0.97656
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 18
## Log-likelihood: -2.03e+04 on 27 Df
```

7.2 Diagnostic plots

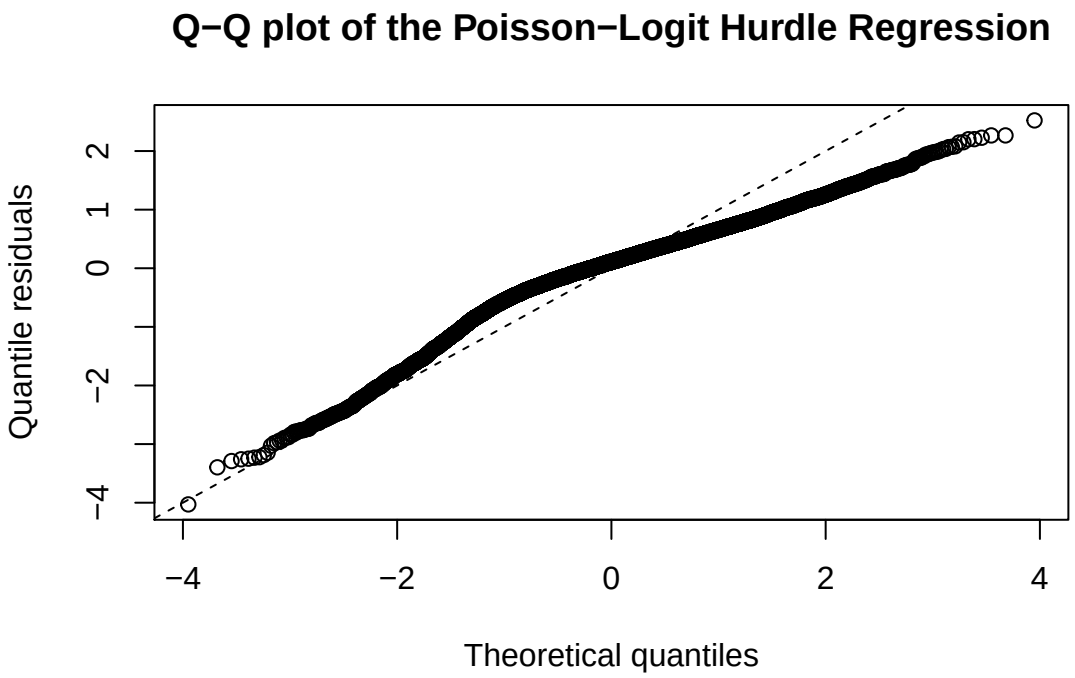
Rootogram as an improved approach to the assessment of fit of a count regression model. Expected counts, given the model, are shown by the thick red line, and observed counts are shown as bars, which in a hanging rootogram are show hanging from the red line of expected count.



As a final check we can also look at the Q-Q plot of the quantile residuals in the hurdle model. These look fairly normal and show no suspicious departures from the model.

Table 5: TARGET Predictions

	n	min	mean	median	max	sd
X1	2523	0.11	3	3	6.8	1.3



7.3 Prediction

We ran predictions on our final model and plotted the distribution next to the distribution from our target in the training data set to compare.

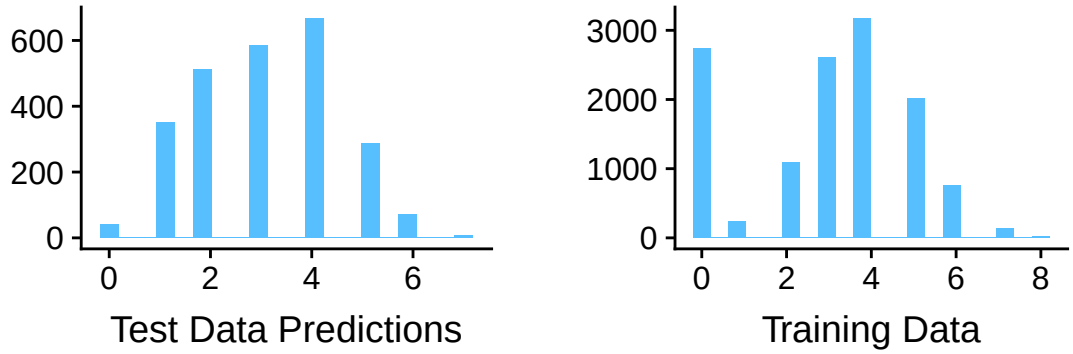


Figure 10: Predictions vs. training data

8 Appendix

The appendix is available as script.R file in `project5_wine` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining